UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

MONIQUE DA SILVA MOORE, MARYELLEN O'DONOHUE, LAURIE MAYERS, HEATHER PIERCE, and KATHERINE WILKINSON on behalf of themselves and all others similarly situated,	)	**Civ No. 11-CV-1279 (ALC) (AJP)**
PLAINTIFFS,	)	
v.	)	
PUBLICIS GROUPE SA and MSLGROUP,	)	
DEFENDANTS.	)	

DECLARATION OF PAUL J. NEALE IN SUPPORT OF PLAINTIFFS' RULE 72(a) OBJECTIONS TO MAGISTRATE JUDGE PECK'S FEBRUARY 8, 2012 DISCOVERY RULINGS

I, Paul J. Neale, declare as follows:

1. I am the Chief Executive Officer and a Managing Director of DOAR Litigation Consulting LLC and have been retained by Sanford Wittels and Heisler, LLP as a consultant and expert in the above-captioned matter.

2. I hold a Bachelor of Arts degree in criminal justice from Temple University.

3. I have advised lawyers and their clients on the management of information in litigation for over 20 years and am a nationally recognized expert on issues relating to the management and production of electronically stored information ("ESI").

4. I am a frequent author, lecturer and CLE instructor regarding the proper management of ESI and on the evolving state of the law and technology as they relate to ESI issues.

5. As a Managing Director at DOAR Litigation Consulting, I am routinely called upon to render expert advice and provide expert testimony on behalf of clients on discovery issues such as ESI preservation, spoliation, cost-shifting, reasonableness, inaccessibility determinations, ESI sanctions and the use of alternative technologies in the analysis and review of ESI.

6. I have appeared as an ESI expert in both state and federal courts and appeared before the rules committee that drafted Federal Rule of Evidence 502 regarding privilege. My testimony to the committee led directly to the Rule 502 Advisory Committee note that states, “a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken 'reasonable steps' to prevent inadvertent disclosure.”

7. I have advised clients on the selection and use of companies that provide predictive coding (or technology assisted review) services.

8. I am thoroughly familiar with the issues related to electronic discovery and ESI in the employment discrimination matter at hand. I have participated in many of the meet-and-confer calls with the defendants and physically appeared on behalf of the plaintiffs in two of the ESI hearings before Magistrate Judge Peck on January 4, 2012, and February 8, 2012.

9. During the hearing on February 8, 2012, I provided testimony regarding the ineffectiveness of defendants’ proposed use of predictive coding with a particular emphasis on the lack of measurability of the accuracy of defendants’ process. Specifically, I described why the use of the defendants’ proposed methods for testing the accuracy of their predictive coding process are insufficient as to allow plaintiffs and the court to effectively measure the accuracy of the results.

10. While I remain in disagreement with the validity of several aspects of the defendants’ predictive coding methodology (e.g. the strong bias toward highly relevant documents to seed the system; and the limited, iterative training focusing solely on relevant documents) as discussed during the ESI hearings, the ultimate, critical flaw in defendants’ protocol regarding the use of predictive coding is the lack of specific measurements of accuracy of their process.

THE USE OF PREDICTIVE CODING REQUIRES MEASURABILITY OF ACCURACY

11. The ability to specifically and defensibly measure the accuracy of any predictive coding (or technology assisted review) methodology is a critical component of the process in order for both parties to agree to its use. This is supported by many academic, judicial and market sources.

12. A measurement of a predictive coding methodology’s accuracy provides the requesting party with the information it needs to determine whether or not the discovery requests have in fact been fulfilled by the producing party. The proper measurement of a predictive coding methodology’s accuracy will be provided only if the protocol that defines the process includes a valid and precise measurement of the accuracy of the review effort responsible for the production.

13. The use of predictive coding requires a precise measurement of recall and precision in order to demonstrate its effectiveness.

Recall

14. Recall is a formulaic calculation of how many instances of any particular value were found as compared to the total number of actual values in the total population. More specifically, how many documents did the system identify as responsive as compared to the total number of actually responsive documents in the entire corpus.

15. The total number of actually responsive documents in the entire corpus is estimated by determining the yield resulting from a review of a statistically valid random sample. In other words, yield is the percentage of responsive documents found as a result of the manual review by a senior attorney of a sample which is applied to the entire corpus of documents. Therefore, if 1,000 documents within a 10,000 document sample are responsive, then the estimated yield would be 10% of the total number of documents in the corpus.

16. The formula for recall is as follows:

$$\textit{Recall} = \frac{\textit{Number of Documents Predicted to Be Responsive}}{\textit{Total Number of Actually Responsive Documents}}$$

17. Therefore, if the system predicts there to be 100,000 documents but the actual number of responsive documents (the yield) is determined to be 125,000 then recall is measured at 80%.

Precision

18. Precision is the measure of how accurately the predictive coding process is determining a particular value. More specifically, how many documents indicated by the system as responsive were determined to be actually responsive.

19. Precision is measured by reviewing a statistically valid sample of the documents that the predictive coding process predicts to be responsive and having a senior attorney manually review those documents to determine how many of those documents are actually responsive.

20. The formula for precision as follows:

$$\textit{Precision} = \frac{\textit{Number of Actually Responsive Documents}}{\textit{Number of Documents Predicted to Be Responsive}}$$

21. Therefore, if a sample of 1,000 documents predicted as responsive is reviewed by the attorney, who determines that 300 of those documents are actually not responsive then precision is measured at 70%.

F-measure

22. F-measure (also known as F_1) is the combination of precision and recall that is often used as a final summary measure for the entire system. The higher the score, the better the performance.

23. The formula is as follows:

$$F_1 = 2 \bullet \frac{Precision \bullet Recall}{Precision + Recall}$$

24. Using the aforementioned examples, the F_1score of 80% recall and 70% precision would be 75%.

GENERALLY ACCEPTED INDUSTRY STANDARDS REQUIRE PREDICTIVE CODING TO INCLUDE A MEASURE OF ACCURACY

TREC

25. The Text Retrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.[1]

26. The TREC process incorporates Recall and Precision as a common evaluation measure to evaluate the quality of a set of retrieved documents and incorporates it into all results and findings.[2]

The Sedona Conference

27. The Sedona Conference is non-partisan research and educational institute dedicated to the advancement of law and policy in the areas of antitrust law, complex litigation, and intellectual property rights. It exists to allow leading jurists, lawyers, experts, academics and others to come together - in conferences and mini-think tanks (Working Groups) - and engage in true dialogue, not debate, all in an effort to move the law forward in a reasoned and just way.[3]

28. In the "Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery", one of the solutions calls for an agreement to

[1] http://trec.nist.gov/overview.html

[2] http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf, http://trec.nist.gov/pubs/trec20/appendices/measures.pdf

[3] http://www.thesedonaconference.org

measure and evaluate the quality of the search and retrieval process and suggests the use of metrics that are currently used in information science, namely, “precision” and “recall.”[4]

29. The “Commentary on Achieving Quality in the E-Discovery Process” indicates that a party must be prepared “to demonstrate to opposing parties, courts, and government agencies, that its chosen method and tool accurately captured a reasonably sufficient number of the relevant, nonprivileged [ESI] in existence, and that the remaining unreviewed and unproduced ESI is irrelevant.” It goes on to make references to the TREC Legal Track research and the use of recall and precision.[5]

30. The Sedona Conference has also put forth their endorsement of TREC Legal Track in an open letter to law firms and companies in the legal tech sector.[6]

Recommind Marketing Literature

31. Recommind’s own marketing literature makes reference to their participation in the 2011 TREC Legal Track program (*see* Recommind, Inc., *2011 TREC Legal Track FAQs*, attached as Exhibit 1) which along with their Predictive Coding Process as outlined in their DESI IV Position Paper (*see* Howard Sklar, Senior Counsel Recommind, Inc., *Using Built-In Sampling to Overcome Defensibility Concerns with Computer-Expedited Review*, attached as Exhibit 2) clearly spells out their agreement for the need to measure accuracy through a calculation of precision, recall and F1 score

Magistrate Judge Andrew Peck

32. Magistrate Judge Peck has been a strong proponent for the use of computer-assisted review and noted in his recent article, “Search, Forward: Time for Computer-AssistedCoding”, that “The object of search is to produce high recall and high precision. Recall is the fraction of relevant documents identified during a review, i.e., a measure of completeness. Precision is the fraction of identified documents that are relevant, i.e., it is a measure of accuracy or correctness.” He goes on to explain that “… if the use of predictive coding is challenged in a case before me, I will want to know what was done and why that produced defensible results. I may be less interested in the science behind the "black box" of the vendor's software than in whether it produced responsive documents with ***reasonably high recall and high precision***.”[7](emphasis added)

[4] http://www.thesedonaconference.org/dltForm?did=Best_Practices_Retrieval_Methods___revised_cover_and_preface.pdf (p194, 205-207, 215)

[5] http://www.thesedonaconference.org/dltForm?did=Achieving_Quality.pdf (p18)

[6] http://www.thesedonaconference.org/content/miscFiles/TREC_OPEN_Letter.pdf

[7] http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202516530534

DEFENDANTS' PROTOCOL IS FATALLY FLAWED BECAUSE IT FAILS TO MEASURE ACCURACY

33. Several essential elements are <u>necessary</u> so that a protocol may permit the measurement of recall and precision. A protocol that will provide valid and actionable measures of the accuracy of a production will include three essential elements.

    (A) An agreed-upon standard of relevance that is transparent and accessible to all parties.
    (B) Measures of accuracy that answer the key questions the requesting party will have about a review for production.
        i) What is the protocol's **recall**? In other words, to what extent has the review found all that I requested?
        ii) What is the protocol's **precision**? In other words, to what extent has the review succeeded in capturing only what I requested?
    (C) An agreed-upon standard of acceptance (e.g., in order to be accepted, a production will have achieved recall and precision scores that meet or exceed some meaningful reference point).

34. In the current protocol, defendants have proposed an evaluation protocol that is deficient with regard to each of these elements and so they cannot provide a meaningful measure of the accuracy of their production (and so that it cannot provide plaintiffs with the information needed to determine whether or not their discovery requests have in fact been fulfilled). In the following, we examine how defendants' protocol stands with regard to each of these elements.

(A) <u>Defendants' Protocol Fails to Define a Standard of Relevance.</u>

35. The foundation for a meaningful measurement protocol requires the development of a well-specified standard of relevance that is documented and accessible to all parties. Without such a governing definition of relevance, there will be no stable standard by which to make the assessments that are the input to any measurement calculation.

36. The plaintiffs have suggested specific language that clearly defines each category of relevance but the defendants have argued against the need for such information. The protocol does discuss a mechanism for resolving disputes over assessments, on a case-by-case basis, in the course of the development of the training set to be used as input to the defendants' chosen review method. Such case-by-case resolution of disputes, however, falls far short of the development of the comprehensive, stable, and well-documented definition of relevance that is needed in order to gather the inputs to meaningful calculations of accuracy and avoid the need to disrupt the process and/or seek assistance from the court.

37. The formulas used for calculating recall and precision, and therefore for calculating accuracy, require an initial assessment of the number of documents that are, in fact, relevant. Without a well-specified standard of relevance that is documented in advance and available to all parties, such calculations will not be meaningful nor replicable by any third-party investigator or auditor. The pre-condition, therefore, for meaningful and replicable measures of accuracy is a well-documented definition of relevance.

(B) Defendants' Protocol Fails to Set Forth How It Will Measure Accuracy

38. As noted above, the two measures of accuracy that are essential to an evaluation of the adequacy of a document production are recall and precision.

39. The current protocol twice uses the words "recall" and "precision," but never discusses how the protocol will calculate these measures, or how the protocol will use any estimates of the measures. Without a defined approach as to how the parties will obtain these estimates of accuracy, it will be impossible for plaintiffs and the court to assess whether the protocol will in fact result in an adequate production.

40. Mere mention, without any further discussion, of the metrics that are at the heart of any meaningful evaluation is symptomatic of the incompleteness of the proposed protocol.

41. Researchers in the field of information retrieval have developed a number of cost-effective and accurate approaches to estimating the recall and precision achieved in a retrieval effort. Any adequate description of an evaluation protocol will contain a detailed discussion of the specific approach it proposes to be followed in obtaining estimates of the accuracy metrics at the heart of the exercise; without such a description, it will be impossible for parties to assess whether the protocol will in fact be able to provide accurate and precise estimates of the essential metrics.

42. To the extent that the defendants are using "sampling" to obtain estimates of recall and precision, these efforts fail, because the protocol does not explain how the use of sampling will be applied to the calculation of recall and precision. For this reason alone, defendants' protocol does not meet generally accepted standards for measuring accuracy.

43. Despite this lack of definition in its sampling approach, I have attempted to anticipate what approach the protocol is suggesting. In my opinion, the defendants' protocol suggests a "yield-based" approach when it first mentions sampling, and an "acceptance-test" approach the second time it mentions sampling. For the reasons I will set forth below, the manner in which the defendants' protocol implements both of these approaches is inadequate to measure recall and precision. Therefore, neither the plaintiffs nor the Court will

be able to determine whether the defendants' protocol has, in fact, retrieved an acceptable level of relevant documents.

Yield-Based Approach

44. The first time the current protocol discusses sampling is in relation to the initial random sample. In this instance, the protocol seems to suggest a "yield-based" approach to the estimation of recall.

45. If suggesting a yield-based approach, then it does not meet an acceptable standard for two reasons: first, the protocol fails to describe the approach in sufficient detail; and second, the initial sample is not representative of all the documents nor is it large enough to provide a precise calculation of recall.

46. With regard to description, an adequate description of a yield-based approach to the estimation of recall would include, at a minimum, the following:
    - A description of the sampling design to be followed in obtaining a yield estimate;
    - A description of the formula for calculating the variance of the yield estimate;
    - A description of the formula for arriving at a 95% confidence interval associated with the yield estimate;
    - A description of the sampling design to be followed in obtaining an estimate of the number of actually relevant documents captured by the retrieval effort;
    - A description of the formula for calculating the variance of the capture estimate;
    - A description of the formula for arriving at a 95% confidence interval associated with the capture estimate;
    - A description of the formula for combining the yield and capture estimates in order to arrive at a recall estimate;
    - A description of the formula for calculating the variance of the recall estimate;
    - A description of the formula for arriving at a 95% confidence interval associated with the recall estimate; and
    - A detailed discussion of the width of confidence intervals for the recall estimate that would be entailed under various conditions (yields and sample sizes).

47. With the exception of the first item noted above, the protocol is lacking in any of the specifics just noted, making it impossible for plaintiffs to assess the adequacy of defendants' proposed approach to the estimation of recall.

48. In terms of substance there are two problems. First, the initial random sample has already been selected and reviewed, prior to all of the data being loaded into the system. In all likelihood, the documents that were included in this initial review are not fully representative of the entire population to be reviewed. As a result, the assessments made at that initial stage will be inconsistent with the selection of

subsequent suggested relevant documents later in the process that are used to train the system.

49. The second substantive problem is that the initial sample would be too small to obtain a reasonably precise estimate of recall under many scenarios. This is because the protocol appears to use the estimate of yield in order to attempt to obtain an estimate of recall. In many cases, especially cases in which the yield of relevant documents is low (as purported to be by the defendants in this case[8]), the sample size required to obtain a precise recall estimate will be many times that required to arrive at a precise yield estimate.

50. In layman's terms, using the sample size for the yield estimate in this case to also determine the recall estimate will result in an unacceptably large confidence interval as detailed below which would possibly result in as much as <u>60% of the responsive documents not being produced</u>.

51. Here is a concrete application of these calculations:

    a. Suppose, for example, we had a population of 3.2 million documents; suppose, then, that, for purposes of estimating the yield of relevant documents, we drew a random sample of 2,399 documents from the population (as is suggested in defendants' protocol). Suppose, further, that, upon review, the sample was found to contain 36 relevant documents. On the basis of that result, we would estimate the full population yield of relevant documents to be 1.5% of the full population, with a 95% confidence interval running from 1.01% to 1.99%. Clearly the sample succeeded in providing us with a precise estimate of the yield.

    b. Suppose, next, we wanted to obtain an estimate of precision. Suppose, in this case, we applied defendants' search methodology to the full population and identified 35,000 documents as relevant. Suppose we then drew a 900-document random sample of the documents so deemed relevant; upon review, we find that 720 of the sampled documents are actually relevant. That result brings us to a precision estimate of 80%, with a 95% confidence interval running from 77.4% to 82.6%. Again, our sampling has enabled us to reach a very precise estimate.

    c. Suppose, finally, that we wanted to combine the results of the two sampling exercises to arrive at a recall estimate. Using the results of our yield sampling, we estimate that the full-population yield is 48,020 documents; using the results of our precision sampling, we estimate that our search methodology succeeded in capturing 28,000 of those relevant documents; we thus arrive at a recall estimate of 58.3%. When, however, we calculate the 95% confidence interval associated with the recall

[8] Mr. Anders had indicated the that yield estimate was 1.5% (see 1/4/2012 transcript at 46:3)

> estimate, we find that it is unacceptably large, running from 39.3% to 77.3%. The sampling we have done, while indeed providing a precise yield estimate, still does not provide a precise recall estimate; it still does not enable us to say, therefore, whether the search effort has been effective or not (it could have captured over 77% of the relevant documents or it could have captured less than 40% of the relevant documents). We would need to have used a much larger yield sample if we wished to arrive at a precise recall estimate.

52. By focusing only on the yield estimate, and not detailing how they would arrive at a recall estimate, defendants have failed to demonstrate that their proposed sample size will enable parties to obtain precise estimates of recall.

Acceptance-Test Approach

53. The second time the protocol mentions sampling is when discussing the random sample of deemed-irrelevant documents during the Quality Control phase of the process. In this instance, the protocol appears to suggest an acceptance-test approach.

54. If it is an acceptance-test approach, then again it does not meet an acceptable standard for two reasons: first, the protocol fails to describe its planned approach in sufficient detail; and second, the sample is not large enough to provide a precise calculation of recall.

55. In terms of description, an effective acceptance-test for ascertaining minimum levels of recall would include the following elements:
    - A description of the sampling design to be followed in obtaining an estimate of the number of actually relevant documents captured by the retrieval effort;
    - A description of the formula for calculating the variance of the capture estimate;
    - A description of the formula for arriving at a 95% confidence interval associated with the capture estimate;
    - Specification of a level of recall at which parties would like a retrieval effort to pass;
    - Specification of the probability with which the retrieval effort should pass when it is at or above the above threshold;
    - Specification of a level of recall at which parties would like a retrieval effort to fail;
    - Specification of the probability with which the retrieval effort should fail when it is at or below the above threshold;
    - Description of the formula for translating the above-noted threshold levels of recall to rates of relevant documents in the subset of documents a retrieval effort has deemed not relevant.
    - Specification of a sampling design (sample size and acceptance threshold) that meets the above requirements.

56. Defendants' protocol <u>fails to specify any of the elements just noted</u>, making it impossible for plaintiffs to assess the adequacy of the proposed approach to the determination of recall.

57. The second problem is that the sample would be too small to obtain a reasonably precise estimate of recall under many scenarios, and indicates that defendants' plans are not well thought out. Defendants propose drawing a sample of 2,399 documents from the subset of documents that have been deemed to be not relevant. For a well-designed acceptance test, this sample size would, in many scenarios, be inadequate so as to create for a greater margin of error.

58. Here is concrete application of these calculations:

    a. Suppose, for example, we had population of 3.2 million documents. Suppose, also, that we had applied our search methodology to the population and identified 1.5% of the population as relevant; suppose, moreover, that, as a result of sampling from the set so identified as relevant, we could assume that our precision was approximately 80%. Suppose, finally, that we wanted to design an acceptance test that would pass 95% of the time when our recall was 90% or higher but fail 95% of the time when our recall was 80% or lower.

    b. In order to meet these requirements we would draw, from the set of documents that the search methodology had identified as not relevant, a <u>sample of 9,289 documents</u>; we would pass the test if, in the sample, we found 19 or fewer relevant documents; we would fail the test if we found 20 or more relevant documents in the sample. A well-designed acceptance test will, in many cases, require larger samples than those suggested in defendants' protocol.

59. In summary, with regard to the estimation of recall and precision, defendants have failed to provide sufficient specifics as to their proposed method to allow plaintiffs to assess whether the method will in fact yield accurate and precise estimates of the measures of interest.

60. To the extent that defendants do suggest certain approaches through sampling, that information does not provide confidence that the suggested method will enable the precise estimation of recall and precision.

61. Finally, it should be noted that the concerns just expressed about the proposed measurement protocols (insofar as they can be gathered from defendants' description of their proposal) apply when estimating recall at the aggregate level (i.e., combining all issue codes into a single "responsive" category).

62. To the extent that it is important to measure recall at the level of individual issue codes, the concerns expressed above areall the more salient (as the problems

already noted will only be exacerbated (due to the lower yields of the specific issue codes).

(C)Defendants' Protocol Fails to Specify a Standard of Acceptance.

63. Finally, an adequate protocol will specify, in advance, a standard of acceptance. Whether that standard is expressed in terms of absolute levels of recall and precision or in terms of relative levels of recall and precision (e.g., the level that could be expected by a viable alternative approach) is a decision that should be made by the parties in advance.

64. A protocol that fails to specify a standard of acceptance in advance sets the stage for further disputes and acrimony. As a result, the implications of any measures of accuracy will remain open to interpretation and dispute.

I declare under the penalties of perjury that the foregoing is true and accurate to the best of my knowledge and belief.

Dated: February 22, 2012

Paul J. Neale